Labour Economics 58 (2019) 21–36

Contents lists available at ScienceDirect

Labour Economics

journal homepage: www.elsevier.com/locate/labeco

# Teaching assistants, computers and classroom management

Helen Johnson [a], Sandra McNally [b,*], Heather Rolfe [d], Jenifer Ruiz-Valenzuela [c], Robert Savage [e], Janet Vousden [a], Clare Wood [f]

[a] Coventry University, Priory Street, Coventry CV1 5FB, UK
[b] University of Surrey and Centre for Economic Performance, London School of Economics, Guildford GU2 7XH, UK
[c] Centre for Economic Performance, London School of Economics
[d] National Institute of Economic and Social Research, 2 Dean Trench Street, Smith Square, London SW1P 3HE, UK
[e] Psycology and Human Development, University College London, 25 Woburn square, London WC1HOAA, UK
[f] Psychology Department, Nottingham Trent University, 50 Shakespeare Street, Nottingham NG1 4FQ, UK

ARTICLE INFO

*JEL Classifications:*
I21

*Keyword:*
Literacy
ICT
Teaching assistants

ABSTRACT

Many students still leave school without a good grasp of basic literacy, despite the negative implications for future educational and labour market outcomes. We evaluate how resources may be used within classrooms to reinforce the teaching of literacy. Specifically, teaching assistants are trained to deliver a tightly structured package of materials to groups of young children aged 5–6. The training is randomly allocated between and within schools. Within schools, teaching assistants are randomly assigned to receive training in either computer-aided instruction or the paper equivalent. Both interventions have a short-term impact on children's reading scores, although the effect is bigger for the paper intervention and more enduring in the subsequent year. This paper shows how teaching assistants can be used to better effect within schools, and at a low cost.

## 1. Introduction

A significant number of children leave primary school with low levels of literacy. Despite much effort to improve basic skills in England, about 11% of children still leave primary school without having achieved the 'expected level' set out in the National Curriculum. This is a long-standing problem in England as it is in many other developed countries. According to an international OECD study, about a fifth of adults in England have low levels of literacy and the problem has not improved amongst young adults compared to older generations (unlike most other countries).[1] The potential implications include lower subsequent educational performance and poor labour market outcomes (e.g. see Vignoles 2016).

There is a large body of evidence showing that teacher quality matters and a small but growing literature showing how interventions can boost teachers' skills (e.g. Taylor and Tyler, 2012).[2] Less is known about the effect of teaching assistants on student outcomes, even though they are used in almost all primary schools in England. In fact, teaching assistants account for about 18% of the average school budget in English primary schools.[3] They usually do not have high-level qualifications and are often used in classrooms to help students with special needs or from low-income backgrounds. Studies about their effectiveness are mostly correlational.[4] In this paper, we evaluate how teaching assistants might be used to better effect the literacy outcomes of young children. The intervention is not to replace core literacy instruction, nor to substantially affect the actual resources available to schools.

The context of the study is a carefully designed programme of small group tuition for 5 year-old pupils in English schools. This has been developed by a team of UK educational psychologists as a balanced, structured reading program that contains a systematic phonics aspect, in line with recommendations in the UK and other English speaking countries. The programme can be delivered in an ICT form (*ABRA-*

* Corresponding author.
*E-mail addresses:* ab6966@coventry.ac.uk (H. Johnson), s.mcnally1@lse.ac.uk, s.mcnally@surrey.ac.uk (S. McNally), H.Rolfe@niesr.ac.uk (H. Rolfe), Robert.savage@ucl.ac.uk (R. Savage), janet.vousden@coventry.ac.uk (J. Vousden), clare.wood@ntu.ac.uk (C. Wood).

[1] OECD PIAAC study, analysed by Kuczera et al. (2016).

[2] Examples of studies showing the importance of teacher quality include Aaronson et al., 2007; Araujo et al., 2016; Chetty et al., 2014a, 2014b; Hanushek et al., 2005.

[3] Times Education Supplement. 2 February 2018. https://www.tes.com/news/exclusive-army-teaching-assistants-continued-expand-even-funding-squeeze-began

[4] The Education Endowment Fund has an evidence summary about TAs. One of the references to how they may be effectively deployed refers to this study, which they commissioned. https://educationendowmentfoundation.org.uk/evidence-summaries/teaching-learning-toolkit/teaching-assistants/

https://doi.org/10.1016/j.labeco.2019.02.006
Received 31 October 2018; Received in revised form 13 February 2019; Accepted 20 February 2019
Available online 23 February 2019

*CADABRA* or *ABRA*), which is widely used in Canada and North America (Abrami et al., 2010), or in a more traditional paper form (Non-ICT).[5] The underlying pedagogy is based on four decades of scientific psychological theory and evidence from a series of meta-analyses of 'what works' in literacy.[6] The core part of this intervention is the training of teaching assistants who are already employed by the school and then the implementation of the small group teaching (which takes place outside of core literacy classes). Specifically, pupils are put together in small groups (3 to 4 pupils) and receive 15 min of teaching four times per week over 20 weeks. Importantly, the intervention does not increase instruction time (i.e. selected pupils receive the treatment while the control group receives 'business as usual' non-core literacy instruction). We can think of this intervention as measuring the effectiveness of redeploying resources within a school rather than the provision of new resources. What is being manipulated is how teaching assistants are being used for a particular year group, holding teacher quality (and the number of teaching assistants employed) constant.

The study is conducted as a Randomised Control Trial. Schools are randomly assigned to receive the treatment. Within treated schools, pupils are randomly assigned amongst three conditions: ICT program (*ABRA*); Non-ICT program (paper equivalent of *ABRA*) and a control group. Within treatment schools, teaching assistants are also randomly assigned to receive training in the ICT and Non-ICT condition and therefore to teach students in one or other group within their school. This design enables us to distinguish between the effects of the underlying pedagogy (common to both) and the effects of the mode of intervention (technology or paper-based). It also enables us to observe whether spillovers occur within treated schools by comparing results with different control groups (i.e. pupils not receiving the treatment in treated schools; pupils not receiving the treatment because they are in control schools). We consider the effects of the intervention at the end of the school year in which it was implemented and also one year later.

Our results show a large initial effect of the program, which is higher for the Non-ICT intervention ($0.18\sigma$ and $0.27\sigma$ for the ICT and Non-ICT interventions respectively).[7] One year later, there is substantial fade-out of effects for pupils assigned to either the ICT or Non-ICT intervention, although the magnitude of this fade-out is in line with other education interventions (e.g. the fade-out for Project Star, as reported by Whitmore Schanzenbach, 2007). The point estimates suggest an effect of about one-third of the initial effect (in either case). There is a significant effect for the Non-ICT treatment if one considers administrative measures of performance the following year.[8] Pupils assigned to the Non-ICT treatment are more likely to achieve the 'expected level' in reading by 6 percentage points (which may be compared to a mean of 74% in the control group). There are also effects for writing and a smaller (but insignificant) effect for maths one year after the end of the intervention. Given the low cost of the intervention, effects of the magnitude presented here are likely to be cost-effective.

Although there is a spillover effect in the same year of the intervention, this is not evident one year later for any outcome. As TAs are with classes at other times of the school day, the most plausible explanation is that the TA is better able to do his/her job generally, thus affecting all students. This study shows how Teaching Assistants might be used within schools to improve the educational outcomes of young people. It also contributes to the literature that gets inside the 'black box' of what is happening inside the classroom.

[5] More specifically, ABRA provides a balanced suite of online activities (alphabetics, fluency, comprehension, and writing) to support reading that can be tailored for context specific purposes.

[6] There is some previous evaluation support based on smaller scale studies (see Section 2).

[7] However, this difference is only statistically significant at the 10 percent level.

[8] This is part of the formal National Curriculum for all children. Key Stage 1 assessments take place at the end of Year 2, when children are aged 7.

The rest of the paper is structured as follows. In Section 2, we give a brief overview of relevant literature. In Section 3, we describe the intervention in detail and in Section 4 we explain the methodology. In Section 5, we present the results. We discuss potential mechanisms in Section 6 before concluding in Section 7.

## 2. Literacy interventions: what do we know?

There have been efforts in many different countries to change approaches to teaching literacy, both for the benefit of children generally as well as for those who have initial reading difficulties. Slavin et al. (2011) reviews developments over the last 25 years in research, policy and practice relating to programs for elementary-aged children who are struggling to learn to read. For example, 'Reading Recovery', developed in New Zealand in the 1970s is one of the best-known and well-researched programmes, and has been disseminated throughout the English-speaking world. This involves individualised instruction for 30 min a day for 12–20 weeks with a specially trained teacher. In the US, successive administrations have encouraged interventions aimed at struggling readers. For example, in the 1990s, the Clinton administration's 'America Reads' initiative encouraged the creation of programmes for volunteer tutors to work with struggling readers. 'Reading First' was the Bush administration's initiative for children in early years of schooling, focused on high-poverty, low-achieving schools with a particular focus on small group interventions for struggling readers. In the UK, there have been various national initiatives designed to improve literacy for all children, such as the National Literacy Strategy in the 1990s and the change in national policy to recommend 'synthetic phonics' to all primary schools in the 2000s (see for example Machin and McNally (2008) and Machin et al. (2018). In the late 2000s, the UK government has also supported 'Reading Recovery' (described above) for low attaining students.

Slavin et al. (2011) review the considerable body of research amongst educationalists/psychologists that now exists on such reading programmes. Among their findings it is observed that small group tutorials can be effective, but not as effective as one-to-one instruction by teachers or paraprofessionals; teachers are more effective than paraprofessionals and volunteers as tutors; and traditional computer-assisted instruction programs have little impact on reading. This finding on the ineffectiveness of computer-assisted programs chimes well with the studies by economists who have evaluated this. Examples of relatively large-scale studies with a strong methodological design include those by Angrist and Lavy (2002), Rouse et al. (2004), and Berlinski and Busso (2017). These studies find no effect of teaching with ICT on pupil learning. A review by Bulman and Fairlie (2016) finds studies of ICT and computer-aided instruction in schools to produce mixed evidence with a pattern of null results, with notable exceptions of studies of developing countries and computer-aided instruction that target maths rather than language.

However, the fact that computer-aided instruction is often found to have zero effect does not mean this need always be the case. One would expect this to be influenced by the underlying pedagogy, the quality of the research design and the training of teachers/teaching assistants that deliver the intervention; as well as the classroom context.[9] Presumably, the reason why many schools use such programs is because they believe they are effective. The program being evaluated here (*ABRA*)[10] has some support from small efficacy Randomised Control Trials (see, for instance, Comaskey et al. (2009), Savage et al. (2009) and Wolgemuth et al. (2011)) and a bigger effectiveness trial (Savage et al., 2013). Savage et al. (2009) randomly allocated 174 pupils into 3 groups: a synthetic phonics intervention group, an analytic phonics intervention

[9] Some studies suggest that technology does have potential to have a positive impact when implemented appropriately (e.g. Archer et al. 2014).

[10] http://www.concordia.ca/research/learning-performance/tools/learning-toolkit/abracadabra.html

**Table 1**
Content of training.

| Introduction to teaching reading: |
|---|
| • How to use the interventions as a tool to teach children skills to maximise their reading outcomes in the broadest sense |
| • Basic reading skills – decoding, fluency, and comprehension |
| • Why the basic reading skills are important to reading outcomes |
| • Teaching multi-ability groups |
| • Managing behaviour in groups/setting group rules |
| The training on the 20 week intervention: |
| • The length and number of sessions to deliver |
| • The aims of each of the activities and how to deliver them |
| • How to keep records of pupils' progress and attendance |
| • How to set (and track) the level of each activity to match that of the pupils |
| • How to access help on each of the activities (in print for Non-ICT, on the laptop for ICT) |
| • How to access (just in time) support during delivery of the intervention |
| Hands-on practice: |
| • Free time to explore the activities and resources |
| • Group time to deliver/role play individual activities |
| • Group time to deliver/role play a whole session (i.e. 3 or 4 activities) |
| • Structured sessions to feedback experience of delivering sessions and activities |
| • Structured sessions to trouble-shoot and share good practice |

Notes: An in-depth description of the content of both interventions can be found in Appendix A and B in McNally et al. (2016).

group and a classroom control group. The intervention groups were both using the *ABRA* computer program. The authors find that both interventions have a significant impact on literacy. Savage et al. (2013) describe a classroom-level Randomised Control Trial (RCT) with just over 1000 pupils, and where the intervention is performed by teachers, also finding improvements in literacy for treated pupils.[11] Our study differs from Savage et al. (2013) along several dimensions. First, the size of the trial in terms of pupils is doubled. Second, this is the first evaluation that has been conducted by a team of independent researchers. Third, the intervention compares an ICT and Non-ICT version of the same program, which are identical in content and only differ in the mode of delivery. Thus, we are able to assess whether the use of technology (i.e. software with graphics, sounds, and cartoon animations designed to appeal to young children) adds value when applying the same underlying pedagogy in the same context (i.e. teaching assistants, in the same schools, undertaking a paper version of the same program). Finally, and most importantly, the research design in this paper includes a clean control group with pupils in schools that do not receive and do not know about the existence of the web-based program while the intervention is in place. Thus, we have a 'clean' control group that represents 'business as usual' for the treatment schools. As we show, within treated schools, non-treated students are affected in the short-term.

## 3. The intervention

Two literacy interventions are evaluated here and both consist of small group tuition for Year 1 pupils in English schools (i.e. pupils of age 5–6): one uses an ICT program (*ABRA*) and the other is identical (i.e. used materials that replicate the ICT intervention) but without using the computer program to deliver the content. Both methods were reviewed by the same independent expert in advance of this study, and teaching assistants (TAs) were trained in the different approaches by academics who are experts in these areas.[12] Table 1 gives a summary of the topics covered by the training approaches. The reading program consists of a balanced 20-week schedule of 15 min lesson plans, consisting of activities to develop phonics, fluency, and comprehension skills.

The ICT intervention, *ABRA*, is a modular game-based literacy intervention that is fixed in content (new activities cannot be added). The games are linked to a series of electronic texts (mainly 'stories', some non-fiction) suitable for beginner readers. The activities are aimed at phonics, word reading fluency, and text comprehension and there was a 20-week schedule of lessons planned for this study.[13] There are extension activities for some of the tasks within *ABRA*, and these can be found in the 'teacher area' of the website. Full details of the program are described in McNally et al. (2016).

The Non-ICT intervention also covered the same 20-week schedule of lesson plans. The paper activities used materials such as magnetic letters and cards and a series of storybooks. To facilitate a clean comparison between the two delivery methods, the Non-ICT activities (especially developed for this study) were matched to each *ABRA* activity using the same stories, vocabulary items, questions, words and letter sounds in all the activities. Thus, the Non-ICT version was identical in content to the ICT version and only differed in terms of the delivery method.

Training occurred after schools had been randomised to the treatment and control conditions (discussed below) and after baseline testing of students in all schools. After school randomisation, treated schools provided the names of the teaching assistants that would participate in the intervention. TAs were already employed by schools and assigned to classes at the beginning of the academic year, prior to randomisation. The intervention has no implications for the number or quality of TAs assigned to particular classes.

For each school, a TA was assigned randomly to the ICT and Non-ICT condition before the training event.[14] Training within the ICT and Non-ICT condition was closely matched in terms of content but tailored for each specific mode of treatment delivery. Each TA was trained for 1.5 days (in a given approach) prior to the start of the intervention, in groups of 12–13 people. This consisted of a one-day training, 'homework' practice tasks and a further half-day of consolidation training. On average, each TA also received approximately 0.6 days of further post-training 'just-in-time' support from the project team (a mix of in-person, phone, and email support).

[11] The effect size is in the region of 0.3-0.4 standard deviations, which varies by outcome measure.

[12] Professor Robert Slavin (University of York, UK and Johns Hopkins University, Baltimore) reviewed plans for how the teaching assistants were to be trained in the different approaches and made recommendations on how the comparability of the different methods could be improved in advance. The training with the use of *ABRA* was provided by Professor Robert Savage (University College London) and the training with the non-ICT methodology was provided by Professor Morag Stuart (University College London).

[13] There are also activities for writing, but the implementation team chose not to include these in the 20-week schedule.

[14] A small number of big schools had two TAs per condition.

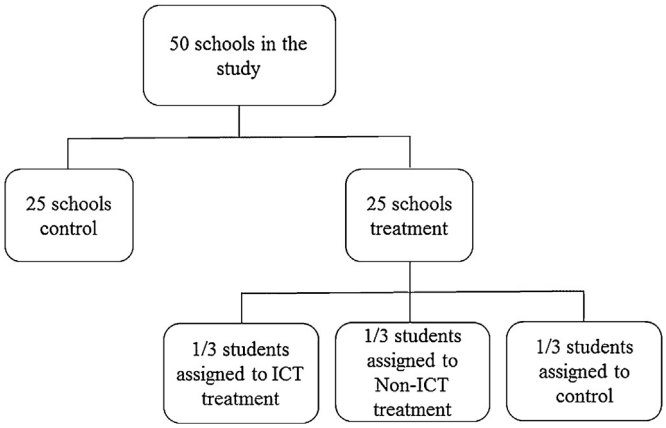


**Fig. 1. Design of the Experiment**.
Notes: The focus of the analysis is on state schools. Within each school, teacher assistants were also randomised to the ICT and Non-ICT condition, respectively.

Both the ICT and Non-ICT TAs received detailed training packs after the training sessions, with a description of the activities and why they were useful. The package included the 20-week plan (available on request) that has guided them on the activities to be performed 4 days per week during the 15-minute sessions. The implementation team at Coventry provided just-in-time support to both groups of TAs on request, and they visited the TAs during the first weeks of treatment to observe how the intervention was delivered and to provide support for the TAs. The TAs were visited again about half way through the intervention.

During training, TAs received a list of pupils assigned randomly to them. Prior to the start of the intervention, TAs had some flexibility in arranging the small groups of pupils (around 3 to 4 pupils per group). The purpose of doing so was to give them the flexibility to divide pupils into appropriate groups, as they normally would do for any other activity. In practice, TAs grouped pupils into groups of 3–4 pupils according to whether they were likely to be able to work well together. This was guided by ability, behaviour, special needs and personality. The process evaluation revealed no issues of concern over implementation or fidelity in delivery. The intervention was found to be well understood by TAs and implemented as intended. This included aspects such as timing, use of materials, and organisation and practical matters. Schools were asked to deliver the programs during literacy-based lessons but not core literacy instruction, including phonics work. This is because the intervention was designed to complement (and not substitute for) normal classroom delivery of literacy (i.e. the intervention did not alter literacy instruction time). The process evaluation suggests this was faithfully adhered to by schools.[15] The broader context of English schools' approach to literacy is very phonics orientated and prescribed (e.g. as discussed in Machin and McNally, 2018). If this intervention is found to benefit children's learning, then this shows that there is value in augmenting standard classroom practice with a wider range of reading activities than are currently used.

## 4. Methodology

The methodology is based on a Randomised Control Trial with two stages: (1) where 50 schools are randomised to treatment and control; (2) where pupils within treated schools are randomly assigned to one of three conditions: ICT, Non-ICT and a control group of students within treated schools.[16] The design of the experiment is illustrated in Fig. 1 and the detail is explained below. An additional layer of randomisation is given by the random assignment of teaching assistants to either the ICT or Non-ICT condition within treated schools.

### 4.1. Participant selection

The implementation team at Coventry University first selected all schools with primary-aged children in the geographical areas near to them, covering schools in the West Midlands.[17] A particular effort was made to encourage schools with disadvantaged intakes to participate during the recruitment stage.[18] The participant schools are those that signed up for the intervention and actually implemented the baseline test for Year 1 students. Randomisation was conducted only after this baseline test had been completed. This applies to 50 schools.[19]

Five schools subsequently dropped out of the intervention, all of them in the treatment group. Of these, three dropped out immediately after randomisation took place and two dropped out later in the year.[20] However, we were able to collect post-intervention data for 4 of these 5 schools that dropped out, and administrative (Key Stage 1 data) is available for all 50 participating schools. This enables us to perform an Intention to Treat (ITT) analysis using most of the original randomised schools, though we also show results that estimate the Treatment on the Treated (TOT).[21] Our full sample consists of 48 schools (or 50 when using the outcome variable from administrative data), half of which were randomly assigned to receive the treatment.[22] Schools were told that they would either receive the treatment in 2014/15 or 2015/16. Thus, the control schools received the treatment in 2015/16. Importantly, the treatment is focused on Year 1 students and thus the cohort of interest to us (i.e. those in Year 1 in 2014/15) will never receive the treatment in control schools.[23] This enables us to consider the effects of the intervention one year later.

### 4.2. Randomisation

School-level randomisation was conducted within pairs of schools. Initially, a number of variables based on administrative data on schools was used to assign each school to its closest pair. These variables included the size of the relevant cohort; the Key Stage 1 average point

[15] More details on the process evaluation can be found in McNally et al (2016).

[16] The trial was registered under the title 'An Evaluation of Teaching Assistant-Based Small Group Support for Literacy' http://www.isrctn.com/ISRCTN18254678. It was conducted according to a protocol set out before the research was conducted. There were only a few small deviations from this protocol that are explained fully in the EEF report (please see McNally et al (2016) and the protocol description here): https://v1.educationendowmentfoundation.org.uk/uploads/pdf/Digital_-_Small_Group_Support_for_Literacy.pdf.

[17] The aim was to recruit about 60 schools, on the basis of power calculations made prior to the evaluation. The calculations to decide on the sample size included in the protocol were performed using the Optimal Design (OD) Software (Spybrook et al, 2011) and is explained further in McNally et al (2016). The implementation team approached all 1682 eligible schools in the West Midlands that included a Year 1 group in the school.

[18] The remit of the commissioner (the Education Endowment Fund) is especially focused on raising the attainment of disadvantaged students.

[19] A further 7 schools originally agreed to take part, but 6 pulled out before baseline testing due to changed circumstances and 1 pulled out after baseline testing (but before randomisation) because they found the process too disruptive.

[20] Two of the schools that dropped out immediately after baseline testing did so because they could not see how to integrate the intervention with their current literacy provision and worried that the children might get confused. One school dropped out during the intervention because of staffing issues and the other because of a change in the head teacher.

[21] Given that we used paired randomisation, we remove from the main analysis both the school for which we did not get any post-test data and its pair (except when the outcomes are defined using Key Stage 1 administrative data, where we can use the full sample of 50 schools).

[22] Results are very similar if we use the 48 schools for all outcome variables.

[23] Furthermore, only 10 of the 25 control schools actually elected to take up the treatment for their Year 1 cohort in 2015/16.

score (i.e. based on teacher assessment for students at age 7) for the relevant cohort in the preceding academic year (2013), and a measure of the percentage of pupils classified as being eligible to receive free school meals.[24] Within each pair, one of the schools was randomly allocated to be in the treatment group, with the other allocated to the control group. We then randomised students in treated schools to one of three groups: (1) the ICT treatment; (2) the Non-ICT treatment and; (3) control pupils in treatment schools.[25] Finally, and as mentioned above, an additional layer of randomisation is given by the random assignment of the teaching assistants participating in the intervention in treated schools, to either the ICT or Non-ICT conditions.

### 4.3. Data and outcome measures

The primary outcome was measured (pre and post-treatment) by the Progress in Reading Assessment (PIRA) test. This is an age-standardised test that evaluates the general reading ability of pupils.[26] Specifically, it assesses reading ability in the following areas: phonics, literal comprehension and reading for meaning, which are the areas that the intervention targets.[27] It has been designed for use at three points in each primary school year (from Reception to Year 6). A separate test is available each term for every year group. It is suitable for whole-class use, with pupils of all abilities. The test booklets are simple and quick to administer (each test takes a maximum of 40 min) and straightforward to mark. The autumn version of the Year 1 PIRA test was used for the baseline test (September 2014, all before randomisation); the summer version of the Year 1 PIRA test was used for the immediate post-treatment testing (July 2015); and the summer version of the Year 2 PIRA test was used for the testing one year after the end of treatment (July 2016).

Assessments were administered by a team of Research Assistants (RAs) employed by Coventry University who did not know to what condition the children had been allocated to. Furthermore, the RAs were blind to the nature of the study – i.e. they were not given any details about the project other than it was a reading project. The baseline PIRA assessment has been scored by Hodder Education. All other tests have been scored (and entered) by a group of RAs hired specifically for this purpose (not those who carried out the assessments), with no knowledge of how schools or pupils have been allocated to the treatment and control groups, and no knowledge of the nature of the project other than it was a reading project.

One year subsequent to the intervention, pupils get to the end of 'Key Stage 1' and receive teacher assessments. The National Curriculum in England is organised around 'Key Stages', within which various goals are made out for children's learning and development and this ends with a formal assessment. Although pupils are assessed by their own teachers at the end of Key Stage 1, there is extensive guidance on how the assessment should be made and it is moderated. As the pupils are in a different school year, the assessment is not made by the same teachers who taught them during the year of this intervention (and there would be no incentive for teachers to manipulate pupil scores on this account – even in the very unlikely scenario that he/she knew who had been in one of the treatment groups in the previous year). The results of the teacher assessment are available in administrative data (the National Pupil Database).

The outcome variables are as follows: (1) PIRA test at endline (i.e., July 2015); (2) PIRA test one year later (July 2016) and (3) Key Stage 1 Reading one year later. The last of these measures is a binary variable, which indicates whether students are at or above the expected level as defined by the National Curriculum. We standardise the PIRA test score to have mean zero and standard deviation of one.[28]

We also incorporate administrative data on pupils as additional control variables: eligibility for free school meals, gender and whether the pupil achieved a good level of development in the Foundation Stage Profile (FSP GLD). The FSP GLD is assessed by teachers when children are at age 5 and in Reception (i.e. their first year of school, which is the year before the intervention takes place) in all schools across the country according to standardised criteria.[29] In this Foundation Stage Profile, pupils are assessed in relation to 17 early learning goals.

The final distribution of pupils in treatment schools before the start of treatment was as follows: ICT treatment (360 pupils), Non-ICT treatment (350 pupils), and control pupils in treatment schools (373 pupils) (see Table A1). There were 1158 pupils in the control schools. Because of school and pupil attrition, our analysis is based on 80 to 95% of the originally randomised sample, depending on the outcome measure analysed (see section below and Table A1 for further details on the level of missing data for the three different outcome variables and across different groups). The slightly higher level of attrition for treated schools shown in Table A1 has to do with the fact that we managed to get endline data for all but one treated school.[30] More details about balance of predetermined characteristics for those observed at endline (for each of the outcome variables) are given in Section 5.

### 4.4. Empirical approach

To estimate the intention-to-treat (ITT) impact, we estimate a regression where the outcome variable is regressed against dummy variables for whether individuals were originally randomised to the ICT or Non-ICT treatment groups (relative to the control group). We also include a dummy for assignment to the control group within treated schools (CT). We control for the school pair in which schools were originally randomised and the baseline test results. We also report results from an augmented regression where we control for predetermined characteristics of students. Given the randomised nature of the intervention, the point estimates should not be greatly affected by the inclusion of additional controls. However, we would expect it to be important for the precision of estimates given a limited number of school clusters. Thus, our most detailed ITT specification can be described as follows:

$$Y_{ist} = \beta_1 ICT_{ist} + \beta_2 NonICT_{ist} + \beta_3 CT_{ist} + \beta_4 Y_{ist-1} + \beta_5 X_{ist-1} + \rho_s + \varepsilon_{ist} \quad (1)$$

Where $Y_{ist}$ is the test outcome for person $i$ in school $s$ at time $t$. As discussed above, we also run this regression using outcomes measured one year later. We are interested in the effects of being assigned to the ICT or Non-ICT treatment (i.e. $\beta_1$ and $\beta_2$) conditional on baseline scores ($Y_{ist-1}$), a vector of personal predetermined characteristics described by $X_{ist-1}$ (which includes gender, eligibility to receive free school meals

[24] In addition, infant schools were paired together (i.e. those catering for pupils of age 4-7; the majority of primary schools cater for pupils of age 4-11).

[25] Note that randomisation is done across the whole year group – even in the case where there is more than one class in a year group. We made an exception for two schools, where we did the randomisation within each class. This is because the classes were in different buildings and the schools would otherwise not have been able to participate in the programme (and would have dropped out after randomisation).

[26] More information on the PIRA test can be found here: https://www.hoddereducation.co.uk/pira. The test provides a wide, thorough coverage at each level within the National Curriculum, from Reception to Year 6. This has been assured by systematically sampling appropriate aspects of the literacy curriculum and Assessing Pupil Progress (APP) in accordance with national guidelines for each year.

[27] The secondary outcomes assess more specific components of reading and are not discussed here (results available on request).

[28] The raw PIRA test score is a continuous variable that can take values from 0 – 25. The age standardised scores range from 70 – 130.

[29] The variable used is a dummy variable that indicates whether the pupil has achieved a good level of development in the Foundation Stage Profile. This is the case if the pupil achieved a level of 2 or 3 in each of COM (Communication), PHY (Physical development), PSE (Personal, Social and Emotional Development), LIT (Language and Literacy) and MAT (Mathematical development) results. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/488745/EYFS_handbook_2016_-_FINAL.pdf.

[30] Moreover, results do not seem to be driven by attrition. Results using KS1 measures (available for all 50 schools) do not change when using the 48 schools for which we have the PIRA test (i.e. the sample available when dropping the school for which we do not have endline test data and its randomisation pair).

**Table 2**
Baseline characteristics—characteristics of treatment and control schools.

| | Control Schools (1) | Treatment schools (2) | *P*-values of the difference in means [Observations] (3) |
|---|---|---|---|
| Total number of teaching assistants (Full-time equivalent) | 12.40 | 12.31 | 0.960 |
| | (6.848) | (7.743) | [50] |
| Total number of teachers (Full-time equivalent) | 15.65 | 16.31 | 0.759 |
| | (6.899) | (10.13) | [50] |
| Ratio of teaching assistants to all teachers | 0.772 | 0.758 | 0.695 |
| | (0.223) | (0.262) | [49] |
| Teachers with Qualified Teacher Status (%) | 97.34 | 98.22 | 0.455 |
| | (4.643) | (3.378) | [50] |
| Mean gross salary of all teachers (in 000 s £) | 36.28 | 35.59 | 0.248 |
| | (1.890) | (2.133) | [50] |
| Size of the Year 1 cohort | 51.44 | 52.76 | 0.712 |
| | (20.02) | (27.33) | [50] |

Notes: Data comes from the School Workforce Dataset (November 2014), except data on the size of the year 1 cohort, that was collected from the implementation team directly from the school records. Columns 1 and 2 show means (first row) and standard deviations (in parentheses). *P*-values are calculated using pairing fixed effects and robust standard errors (column 3). The number of observations is shown in squared brackets in column 3.

prior to treatment and whether the pupil achieved a good level of development in the Foundation Stage Profile), and the school pair $\rho_s$. Standard errors are clustered at the level of the school (i.e. the first stage of randomisation). We are also interested in establishing whether there is any spillover effect of the treatment to control students within treated schools (i.e.$\beta_3$).

We estimate this regression for different subgroups.[31] These subgroups are defined on the basis of free school meal status; gender; above median attainment on pre-test (i.e. PIRA test at baseline). This is of interest in that the effects of the treatment may be heterogeneous between pupils with different characteristics.

Given that 5 schools in the treatment group dropped out (3 immediately after randomisation, and 2 during the intervention), we also estimate Instrumental Variable regressions, using the initial random allocation of students as instruments for the final treatment received. See the 'Note on Methodology' in the Appendix for further detail.

## 5. Results

### 5.1. Balance at baseline

Table 2 shows characteristics of treatment and control schools in terms of the number of teaching assistants (TAs), teachers, the ratio of TAs to teachers, teacher qualifications, salaries and the size of the Year 1 cohort. There is very little numerical difference between those schools assigned to treatment and control in these respects. However, as there are only 50 schools in the sample, any differences are unlikely to be statistically significant. There are about 50 pupils on average within the Year 1 group, which implies about two classes per school. The ratio of TAs to teachers is very close to the national average and close to 0.8 for both treated and control schools. This implies that on average, there is almost one TA per teacher.

Table 3 shows characteristics of TAs within treatment schools that are assigned to the ICT and Non-ICT conditions. The information in Panel A of Table 3 is available for all teaching assistants in treated schools (except for the 3 schools that dropped out immediately after randomisation); and for slightly less TAs in Panel B. As TAs were randomly assigned to the ICT and Non-ICT condition, it is not surprising to see that for the most part, their characteristics are similar on average within each condition. The average TA is in her/his early 40's with about 10 years of experience as a TA.[32] The percentage with qualifications of 'level 3 or more' (corresponding to at least upper secondary education) is 84% for those assigned to the ICT condition and 67% for those assigned to the Non-ICT condition.[33] Information from the TA baseline survey shows that most TAs use information technology (IT) professionally both for the teaching of literacy and numeracy and over 40% use IT professionally every day or for every lesson. For the most part TAs feel comfortable using IT for teaching. This applies to 68% of those TAs assigned to the ICT condition and 47% of TAs assigned to the Non-ICT condition.

Table 4 shows characteristics of students assigned to control and treated schools (columns 1 and 2, respectively); and then within treated schools, those assigned to the ICT, Non-ICT or control condition (columns 3, 4 and 5, respectively). The characteristics are those used in the regression analysis: the student's gender; eligibility for free school meals; whether he/she has achieved a 'good' level of development as measured by teachers in the previous year for the Foundation Stage Profile (described above); and the baseline PIRA reading test. There is almost no difference between the groups with respect to any of these characteristics. The one exception is whether pupils were assessed as having a 'good level of development' within the Foundation Stage Profile.[34] On average, this is higher in control schools (at 54%) compared to treatment schools (at 48%). Otherwise, the groups are fairly well balanced.[35]

We analyse whether attrition is a threat to validity to our estimates by checking balance at endline, for each of the three outcome variables. The results are very similar to those found at baseline and for the three outcomes and are available upon request. Therefore, attrition has not worsened balance on observables across the different conditions. Nonetheless, we show results with and without controlling for detailed baseline characteristics for the main specifications.

### 5.2. Main results for reading

Estimates of the 'Intention to Treat Effects' are shown in Table 5. Columns (1) and (2) show estimates of Eq. (1) for all students. Columns (3) and (4) exclude control students within treatment schools (i.e. only using treated students in treatment schools and all students in control

[31] Having made the point about spillover effects with the overall results, when showing heterogeneous effects, we only report coefficients on the interaction between intervention groups (ICT and Non-ICT) and relevant subgroups. Results are almost identical to excluding the non-treated group of pupils within treatment schools altogether.

[32] Only 3 out of the 52 TAs are male (1 in the ICT and 2 in the Non-ICT condition).

[33] In terms of tertiary education, 28% of TAs in the ICT condition have a Higher Education degree; and 8% of the TAs in the Non-ICT condition.

[34] The p-value is 0.01. There is one other difference where the p-value is less than 0.10 (i.e. 0.09). There are fewer females within the control condition in treated schools compared to the two treatment conditions (i.e. 45% compared to about 51%).

[35] This is also the case if we do the balancing test excluding the school that dropped out of the experiment, for which we could not conduct an endline reading test.

**Table 3**
Characteristics of TAs assigned to each condition.

| | ICT (1) | Non-ICT (2) | *P*-values of the difference in means [Observations] (3) |
|---|---|---|---|
| Panel A. Information from Curriculum Vitae of Teaching Assistants | | | |
| Age TA in first term academic year 2014–2015 | 42.46 | 42.57 | 0.970 |
| | (11.76) | (8.417) | [49] |
| Years of teaching assistant experience | 9.800 | 10.46 | 0.747 |
| | (7.331) | (7.271) | [52] |
| TA has any qualification of level 3 or more | 0.840 | 0.667 | 0.154 |
| | (0.374) | (0.480) | [52] |
| Panel B. Information from baseline surveys | | | |
| Use of IT (professionally) for literacy | 0.955 | 0.868 | 0.336 |
| | (0.213) | (0.347) | [42] |
| Use of IT (professionally) for numeracy | 0.955 | 0.816 | 0.17 |
| | (0.213) | (0.398) | [42] |
| Use IT professionally every day or lesson | 0.409 | 0.457 | 0.769 |
| | (0.503) | (0.513) | [40] |
| TA feels comfortable or very comf. using IT for teaching | 0.682 | 0.474 | 0.185 |
| | (0.477) | (0.512) | [42] |

Notes: The information in this table comes from data collected via standardised curriculum vitae sheets and other pre-information survey. Columns 1 and 2 show means (first row) and standard deviations (in parentheses). *P*-values are calculated using robust standard errors (column 3). [Results are very similar when we also include school fixed effects or when we cluster the standard errors at the school level. Due to the low number of observations and clusters, and the fact that in the second panel we miss information for some of the TAs in some categories, we show the results without including school fixed effects and without clustering standard errors at the school level]. Observations have a weight of 1 if there is only one teaching assistant per group; and 0.5 when there are two teaching assistants per group (due to replacements). The number of observations is shown in squared brackets in column 3.

**Table 4**
Balance checks at baseline: students.

| | Baseline Variable Means and Standard Deviation | | | | | *P*-values of the difference in means [Observations] | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Control Schools (1) | Treatment schools (2) | ICT (3) | Non-ICT (4) | Control in Treatment schools (5) | [2] vs [1] (6) | [3] vs [1] (7) | [4] vs [1] (8) | [4] vs [3] (9) | [5] vs [3] (10) | [5] vs [4] (11) |
| Panel A. Individual characteristics | | | | | | | | | | | |
| Female | 0.498 | 0.494 | 0.516 | 0.513 | 0.455 | 0.555 | 0.466 | 0.677 | 0.963 | 0.087 | 0.106 |
| | (0.500) | (0.500) | (0.500) | (0.501) | (0.499) | [2221] | [1511] | [1501] | [696] | [720] | [710] |
| FSM | 0.216 | 0.229 | 0.219 | 0.232 | 0.236 | 0.527 | 0.665 | 0.587 | 0.779 | 0.8 | 0.952 |
| | (0.411) | (0.420) | (0.414) | (0.423) | (0.425) | [2203] | [1498] | [1486] | [692] | [717] | [705] |
| FSP GLD | 0.543 | 0.482 | 0.482 | 0.500 | 0.466 | 0.010 | 0.057 | 0.27 | 0.605 | 0.633 | 0.381 |
| | (0.498) | (0.500) | (0.500) | (0.501) | (0.500) | [2210] | [1505] | [1492] | [693] | [718] | [705] |
| Panel B. Baseline test | | | | | | | | | | | |
| Std PIRA | 0.0328 | −0.0513 | −0.0510 | −0.0412 | −0.0609 | 0.233 | 0.230 | 0.155 | 0.661 | 0.710 | 0.923 |
| | (1.000) | (0.998) | (1.019) | (0.959) | (1.015) | [2160] | [1464] | [1459] | [677] | [701] | [696] |

Notes: The sample for variables in Panel A includes all available observations in the National Pupil Dataset/survey records. The sample for the variable in Panel B includes all students sitting the baseline PIRA test. The variable in Panel B is standardised using the mean and standard deviation of all available observations at baseline. FSM eligibility: pupil recorded as eligible for free school meals on Census day. FSP GLD: pupil has achieved a good level of development—achieved level of 2 or 3 in each of COM, PHY, PSE, LIT and MAT results. PIRA is the progress in Reading Assessment test, our primary outcome. Standard deviations are in parentheses in columns 1–5 and the available observations for the respective samples are in squared brackets in columns 6–11. P-values are calculated using pairing fixed effects (columns 6–8) and school fixed effects (columns 9–11). Standard errors are clustered at the unit of randomisation: i.e., at the school level in columns 6–8, and at the student level in the within school comparisons (i.e., robust standard errors are used in columns 9–11).

schools). In each case, we show a specification with minimal controls (i.e. the school pair dummies and the baseline reading score) and an augmented version (including controls for gender, eligibility for free school meals and whether the pupil achieved a 'good level of development' in the Foundation Stage Profile at age 5). The simple specification is shown in columns (1) and (3) and the augmented specification is shown in columns (2) and (4). We show three panels of results, with Panel A being the 'intention to treat' effect within the same school year (i.e. about two months after the end of treatment). Panel B shows results when the outcome variable is the PIRA reading test administered one year later.[36] Panel C shows results when the outcome variable is defined as a binary variable indicating whether the student achieves the 'expected level' in the Teacher Assessment that is conducted one year after the intervention (in line with national requirements described above).[37]

In each case, the point estimates of the effects are slightly higher in the augmented specification. Unsurprisingly, the estimated effect of assignment to the ICT and Non-ICT conditions is approximately the same whether or not we exclude control students within treatment schools.

[36] This is the Year 2 Summer version of the test, to take into account that students are a year older.

[37] Note that in each of the specifications, we have used the maximum number of observations available for each outcome. However, reducing the number of observations to include the same observations for each specification and outcome does not change the results. Results are available upon request.

**Table 5**
Intention to treat effects: main results.

| | All students | | Excluding control students in treated schools | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| A. Outcome: PIRA test at *endline* | | | | |
| ICT | 0.144 | 0.179** | 0.150 | 0.186** |
| | (0.087) | (0.079) | (0.090) | (0.081) |
| NONICT | 0.246*** | 0.272*** | 0.259*** | 0.284*** |
| | (0.082) | (0.075) | (0.083) | (0.076) |
| CT | 0.116 | 0.167** | | |
| | (0.082) | (0.074) | | |
| Students | 1901 | 1884 | 1591 | 1576 |
| P value: ICT=NONICT=CT=0 | 0.0142 | 0.0057 | | |
| P value: ICT=NONICT | 0.104 | 0.102 | 0.086 | 0.092 |
| P value: ICT=CT | 0.579 | 0.821 | | |
| P value: NONICT=CT | 0.017 | 0.039 | | |
| B. Outcome: PIRA test at *endline + 1* | | | | |
| ICT | 0.053 | 0.077 | 0.055 | 0.078 |
| | (0.073) | (0.072) | (0.075) | (0.073) |
| NONICT | 0.072 | 0.094 | 0.081 | 0.101 |
| | (0.084) | (0.079) | (0.086) | (0.082) |
| CT | −0.021 | 0.015 | | |
| | (0.078) | (0.073) | | |
| Students | 1799 | 1785 | 1501 | 1488 |
| P value: ICT=NONICT=CT=0 | 0.3286 | 0.3633 | | |
| P value: ICT=NONICT | 0.752 | 0.789 | 0.650 | 0.703 |
| P value: ICT=CT | 0.16 | 0.271 | | |
| P value: NONICT=CT | 0.113 | 0.156 | | |
| C. Outcome: Key Stage 1 Reading at *endline + 1* (at or above the expected reading level) | | | | |
| ICT | 0.008 | 0.019 | 0.008 | 0.018 |
| | (0.025) | (0.025) | (0.024) | (0.025) |
| NONICT | 0.048* | 0.055** | 0.048* | 0.055* |
| | (0.027) | (0.027) | (0.028) | (0.028) |
| CT | −0.021 | −0.006 | | |
| | (0.024) | (0.025) | | |
| Students | 2129 | 2111 | 1770 | 1756 |
| P value: ICT=NONICT=CT=0 | 0.0124 | 0.0526 | | |
| P value: ICT=NONICT | 0.163 | 0.146 | 0.160 | 0.148 |
| P value: ICT=CT | 0.217 | 0.335 | | |
| P value: NONICT=CT | 0.001 | 0.007 | | |
| Mean outcome in control schools | 0.739 | 0.741 | 0.739 | 0.741 |
| *Control variables:* | | | | |
| Baseline PIRA test | √ | √ | √ | √ |
| Gender, FSM, FSP GLD | | √ | | √ |

Notes: Intention to treat estimates. Outcome variables: PIRA test at endline is the standardised score of the PIRA test taken at the end of treatment. PIRA test at endline +1 is the standardised score of the PIRA test taken a year after the end of treatment. KS1 reading at endline + 1 is a dummy variable that equals 1 if the student is at or above the expected reading level at the end of Key Stage 1. ICT and NONICT are the intention to treat dummies. CT is an intention to treat dummy equal to 1 for pupils in the control group of treatment schools. All available students used in columns 1 and 2. In columns 3 and 4, students that were in the control group of treated schools are excluded. All regressions control for randomisation pair dummies. FSM eligibility: pupil recorded as eligible for free school meals on Census day. FSP GLD: pupil has achieved a good level of development—achieved level of 2 or 3 in each of COM, PHY, PSE, LIT and MAT results. Standard errors (in parentheses) are clustered at the school level, with * $p<0.10$; ** $p<0.05$; *** $p<0.01$. Number of schools: Panels A and B (48), Panel C (50).

This is because we include a binary variable for whether or not students are assigned to that group (in columns 1 and 2).

We first consider the short-term effects of the intervention on the reading test conducted at the end of the same school year (Panel A, Table 5). The effect of being assigned to the ICT condition moves from $0.14\sigma$ to $0.18\sigma$ from the simple to the augmented specification. The effect of being assigned to the Non-ICT condition moves from $0.25\sigma$ to $0.27\sigma$. Although not statistically different from each other, the increase in coefficients between the simple and augmented specification may be explained by the fact that there is an imbalance between the treatment and control group (favouring the latter) with regard to the proportion of children with a 'good level of development' the previous year (i.e. according to the Foundation Stage Profile, as explained in Section 4.3).

Both interventions have a significant effect; although the impact of the Non-ICT intervention is about 50% bigger (and the p-value of the difference between assignment to the ICT and Non-ICT intervention is just over 0.10). However, the effect of being assigned to the control condition within treatment schools (captured by the CT dummy in Table 5) is almost the same as being assigned to the ICT condition (and is not significantly different). Thus, there is a substantial spillover effect. As discussed in detail in Section 6, the most likely explanation is that TAs were able to improve how they worked with all the pupils as a result of their training. The TAs were not employed especially for this project.

They were drawn from those already working with Year 1 pupils and did plenty of other literacy activities outside the intervention time. Hence, there would have been opportunity for TAs to use any new skills they had learnt to help pupils informally at other times.

Panels (B) and (C) enable us to consider the effects of the intervention in the next school year. By this time, pupils will have been exposed to another full year of teaching with a different teacher and different teaching assistants. In Panel B, the outcome variable is the PIRA reading test. Any spillover effect disappears as the point estimate is close to zero for being assigned to the control condition within treatment schools. The magnitude of the intention to treat effect of being assigned to the ICT or Non-ICT condition reduces considerably. In the augmented specification, the point estimate is 0.08$\sigma$ and 0.10$\sigma$ for the ICT and Non-ICT condition respectively. However, the standard errors remain roughly the same as in Panel A, which is almost as high as the estimated effects. Thus, at conventional levels of significance, we are unable to say whether or not the intervention continued to have an effect on pupils when using the PIRA test.

In Panel C, we show results where the outcome variable is whether or not the pupil achieved the 'expected reading level' according to the ('Key Stage 1') Teacher Assessment. The baseline (in the control group) is 74 %. Again, there is no evidence of a spillover effect (with the point estimate being close to zero). Estimates of the intention to treat effect are 0.02 and 0.06 (i.e. 2 and 6 percentage points) in the ICT and Non-ICT conditions respectively within the augmented specification. This is significantly different from zero in the case of the Non-ICT condition. Thus, these results give firmer evidence that the effect of the intervention did endure for the Non-ICT condition.

Table A2 shows the impacts of the ICT and Non-ICT conditions when we scale up the results to show the 'Treatment on the Treated' effects. In the augmented specification, point estimates increase slightly to 0.22$\sigma$ and 0.33$\sigma$ when using the PIRA at endline outcome variable for the ICT and Non-ICT conditions, respectively (column 2); to 0.09$\sigma$ and 0.11$\sigma$ one year later (though not statistically significant, column 4); and to 0.02 and 0.07 (i.e. 2 and 7 percentage points) when using the binary variable capturing whether the student has achieved the expected reading level at the end of Key Stage 1 (column 6). The estimated impacts are close to the ITT results because the assignment to treatment and the final treatment received were not very different in most cases (as can be seen by the magnitude of the main coefficients in the ICT, Non-ICT and CT first stages in Panels B, C and D).

It is difficult to compare the reading test to the teacher assessment because the latter is a binary variable and the former is a continuous variable. Of course, they are also different types of assessment and may give different results for that reason. To make results more comparable, we convert the reading test to a binary variable based on how the teacher assessment indicator corresponds to the average reading test score (at endline and endline+1, respectively) within control schools.[38] Results are reported in Table 6. Column (1) shows results where the outcome is the PIRA reading test at the end of the same school year. Columns (2) and (3) show results where the outcome is measured one year later either in the age-adjusted version of the same reading test (column 2) or in the teacher assessment (column 3). Here we report coefficients on the other variables because it is interesting to notice how the magnitudes of the coefficients are similar for the two different assessments measured at the same time (i.e. columns 2 and 3). With regard to the main coefficients of interest, a comparison between columns 2 and 3 shows that results are very similar if we try to measure the reading test and the teacher assessment on a comparable (binary) scale.[39] Comparing point estimates for the outcome variable in the same year as the intervention (column 1) and one year later (columns 2 or 3) suggests that the effect one year later might be around one-third of the original effect.

**Table 6**
Intention to treat effects: binary outcome measures.

| | PIRA dummy (1) | PIRA dummy+1 (2) | Ks1 read endline+1 (3) |
|---|---|---|---|
| ICT | 0.068* | 0.037 | 0.019 |
| | (0.038) | (0.039) | (0.025) |
| NONICT | 0.121*** | 0.043 | 0.055** |
| | (0.039) | (0.035) | (0.027) |
| CT | 0.092** | 0.026 | −0.006 |
| | (0.037) | (0.037) | (0.025) |
| Std PIRA baseline | 0.209*** | 0.184*** | 0.160*** |
| | (0.012) | (0.015) | (0.014) |
| Female | −0.027 | −0.002 | −0.034* |
| | (0.021) | (0.021) | (0.019) |
| FSM | −0.049** | −0.061** | −0.078*** |
| | (0.023) | (0.030) | (0.026) |
| FSP GLD | 0.232*** | 0.166*** | 0.223*** |
| | (0.029) | (0.033) | (0.026) |
| P value: ICT=NONICT | 0.173 | 0.859 | 0.146 |
| Mean outcome in control schools | 0.453 | 0.535 | 0.741 |
| Students | 1884 | 1785 | 2111 |
| Schools | 48 | 48 | 50 |

Notes: Intention to treat estimates. Binary outcome variables: PIRA dummy: equals 1 if the student has a PIRA endline score equal or bigger than the mean PIRA endline score observed for students in control schools working at the KS1 expected reading level. PIRA+1 dummy: equals 1 if the student has a PIRA endline+1 score equal or bigger than the mean PIRA endline+1 score observed for students in control schools working at the KS1 expected reading level. KS1 read at endline + 1 is a dummy variable that equals 1 if the student is at or above the expected reading level at the end of Key Stage 1. ICT and NONICT are the intention to treatment dummies. CT is an intention to treat dummy equal to 1 for pupils in the control group of treatment schools. All regressions control for FSM, female and FSP GLD dummies, standardised baseline PIRA tests, and the randomisation pair dummies. FSM eligibility: pupil recorded as eligible for free school meals on Census day. FSP GLD: pupil has achieved a good level of development—achieved level of 2 or 3 in each of COM, PHY, PSE, LIT and MAT results. Standard errors (in parentheses) are clustered at the school level, with * p<0.10; ** p<0.05; *** p<0.01.

### 5.3. Results for other subjects

Although the intervention was targeted on activities particularly important for reading, it might also impact on other subjects. There is an obvious connection between reading and writing. Machin and McNally (2008) show that there is a strong relationship between reading demands of tests in maths and reading. Specifically, an analysis done on the age 11 reading and maths test showed that the reading demand of the maths test (based on text difficulty) is nearly 70% of what it is in the reading assessment. We do not have test outcomes for other subjects immediately after the intervention but we do have Teacher Assessments for reading, writing and maths in administrative data at the end of the subsequent year when pupils are age 7.

Table 7 shows results for writing and maths respectively where the outcome variable is one if the pupil achieves at least the 'expected level' in these subjects. The effect is only statistically significant in the case of writing and for the Non-ICT treatment only. Specifically, the effect of assignment to the Non-ICT condition increases the probability of achieving the 'expected level' in writing by 0.08 in the augmented specification (i.e. 8 percentage points). The point estimate for maths is also positive (0.05) but not statistically significant. Assignment to the ICT condition does not show effects that are statistically significant. However, point estimates are 0.04 and 0 for writing and maths, respectively, and thus show a pattern of results that is consistent with estimates for the Non-ICT condition, and with the overall short-term results.

[38] We refer the reader to the notes in Table 6 for more detail on how we construct the binary variables at endline and endline+1 (with information from the continuous PIRA at endline and PIRA at endline+1, respectively).

[39] The results are very similar if we use probit/logit regressions for binary outcome variables.

**Table 7**
Results for other subjects, one year later.

| | (1) | (2) |
|---|---|---|
| A. Outcome: Key Stage 1 Writing at *endline + 1* (at or above the expected writing level) | | |
| ICT | 0.028 | 0.040 |
| | (0.032) | (0.032) |
| NONICT | 0.069** | 0.081** |
| | (0.033) | (0.035) |
| CT | −0.019 | 0.002 |
| | (0.037) | (0.035) |
| P value: ICT=NONICT | 0.054 | 0.052 |
| Mean outcome in control schools | 0.619 | 0.620 |
| B. Outcome: Key Stage 1 Maths at *endline + 1* (at or above the expected maths level) | | |
| ICT | −0.009 | 0.003 |
| | (0.032) | (0.031) |
| NONICT | 0.038 | 0.047 |
| | (0.030) | (0.031) |
| CT | −0.008 | 0.004 |
| | (0.031) | (0.031) |
| P value: ICT=NONICT | 0.036 | 0.035 |
| Mean outcome in control schools | 0.712 | 0.713 |
| Students | 2129 | 2111 |
| *Control variables:* | | |
| Baseline PIRA test | √ | √ |
| Gender, FSM, FSP GLD | | √ |

Notes: Intention to treat estimates. Outcome variables: Key Stage 1 Writing (Maths) is a dummy variable that equals 1 if the student is at or above the expected writing (maths) level at the end of Key Stage 1. ICT and NONICT are the intention to treatment dummies. CT is an intention to treat dummy equal to 1 for pupils in the control group of treatment schools. All regressions control for the randomisation pair dummies. Standard errors (in parentheses) are clustered at the school level, with * $p<0.10$; ** $p<0.05$; *** $p<0.01$.

### 5.4. The distribution of test-score gains

It may be that gains vary across the test score distribution. In Table 8, we show results from quantile regressions using the reading test administered at the end of the intervention and one year later. These results show that the Non-ICT intervention has a fairly uniform effect throughout the distribution, except at the 90th percentile (where the point estimate is higher). The point estimate for the ICT intervention is smaller at either extreme (10th or 90th percentile) compared to the middle when the outcome variable is measured at endline (Panel A). One year after the end of the intervention the point estimate for the Non-ICT intervention is also similar (though smaller) through the distribution (Panel B). In contrast, the point estimate for the ICT intervention is bigger at the lower end of the distribution (at 25th percentile and below) compared to at the median and above. However, when running the quantile regressions simultaneously, we can never reject the null hypothesis that test score gains are the same across the distribution.

### 5.5. Heterogeneity

In Table 9, we show results where each treatment dummy is interacted by an individual characteristic: whether the pupil is eligible to receive free school meals (FSM) (panel A); gender (panel B); and whether he/she is above or below the median of the baseline test (panel C). In each case, we include four "treatment" variables defined according to the ICT/Non-ICT treatment status and the characteristic under study. We show three columns of results: the reading test at the end of the intervention year (column 1), the same reading test at the end of the subsequent year (column 2) and a binary variable for whether the pupil achieved the 'expected level' in the Key Stage 1 teacher assessment (also one year after the intervention).

The short-term effect of the intervention was much stronger for FSM pupils compared to non-FSM pupils. For FSM students, the effect was about half of a standard deviation for both the ICT and non-ICT conditions. This would close the gap between FSM and non-FSM students (as this is about $0.30\sigma$ whereas the effect of the Non-ICT intervention was $0.21\sigma$ for non-FSM pupils). The group for whom the intervention was least effective was non-FSM students assigned to the ICT condition (where the point estimate is $0.11\sigma$ and not statistically significant). However, these effects all diminish one year after the intervention. The point estimates suggest that the group least likely to benefit are still the non-FSM students assigned to the ICT condition whereas effects are more likely to endure for FSM students.

In panel B, we show effects by gender. Although point estimates for the short-term effect suggest a slightly bigger effect for girls than boys, the difference is not statistically significant. There is fade-out for all groups. However, the point estimates suggest that girls assigned to the Non-ICT condition benefit most in the short-term (column 1) and also in the longer term if we consider the indicator variable for whether pupils achieve the expected level in reading (column 3). Girls assigned to the Non-ICT condition are more likely to achieve this standard by 9 percentage points whereas the point estimates are smaller and not statistically significant for girls assigned to the ICT condition or for boys assigned to either condition.

Finally, in panel C, we show results according to whether the pupil scored above or below the median of the baseline PIRA test. The first column suggests that the short-term effect of the Non-ICT intervention was about the same, regardless whether the pupil was above or below the median. The magnitude of the effect is also similar to those assigned to the ICT intervention if they scored below the median in the baseline test. A lower point estimate (which is not statistically significant) is found for pupils above the median who were assigned to the ICT intervention. Although these effects fade out in the subsequent year, a similar pattern of effects is observed for the reading test (column 2). The teacher assessment outcome (column 3) shows a similar point estimate for the Non-ICT treatment for pupils above and below the median (though only marginally significant in the case of the former). The point estimate is only slightly lower for above-median pupils exposed to the

**Table 8**
Distributional effects—reading.

| | 0.1Q (1) | 0.25Q (2) | 0.50Q (3) | 0.75Q (4) | 0.90Q (5) |
|---|---|---|---|---|---|
| A. Outcome variables defined at endline (i.e., using PIRA at endline) | | | | | |
| ICT | 0.106 | 0.221** | 0.187** | 0.246** | 0.150 |
| | (0.107) | (0.109) | (0.095) | (0.096) | (0.127) |
| NONICT | 0.239*** | 0.221** | 0.235*** | 0.225** | 0.355** |
| | (0.080) | (0.087) | (0.091) | (0.100) | (0.140) |
| Students | 1884 | 1884 | 1884 | 1884 | 1884 |
| Schools | 48 | 48 | 48 | 48 | 48 |
| P-value Parente-Santos Silva test | 0.001 | 0.000 | 0.000 | 0.000 | 0.105 |
| A. Outcome variables defined at endline+1 (i.e., using PIRA at endline+1) | | | | | |
| ICT | 0.159*** | 0.120 | 0.058 | 0.040 | 0.014 |
| | (0.051) | (0.077) | (0.084) | (0.080) | (0.084) |
| NONICT | 0.105* | 0.097 | 0.120 | 0.095 | 0.066 |
| | (0.055) | (0.077) | (0.083) | (0.120) | (0.079) |
| Students | 1785 | 1785 | 1785 | 1785 | 1785 |
| Schools | 48 | 48 | 48 | 48 | 48 |
| P-value Parente-Santos Silva test | 0.410 | 0.000 | 0.000 | 0.003 | 0.984 |

Notes: Intention to treat estimates. Outcome variables: PIRA test at endline is the standardised score of the PIRA test taken at the end of treatment. PIRA test at endline +1 is the standardised score of the PIRA test taken a year after the end of treatment. ICT and NONICT are the intention to treatment dummies. The CT intention to treat dummy (dummy equal to 1 for pupils in the control group of treatment schools) is included but not shown in the table. All regressions control for FSM and female dummy, FSP GLD, standardised baseline PIRA tests, and the randomisation pairs. We cluster standard errors at the school level in all cases where the Parente–Santos Silva test for intra-cluster correlation rejects the null of no intra-cluster correlation. In the two exceptions where the null is not rejected, we do not cluster by school and use robust standard errors. * p<0.10; **p<0.05; ***p<0.01.

ICT treatment (though not statistically significant) and close to zero for below-median pupils exposed to the ICT treatment.

## 6. Mechanisms

The training of teaching assistants both for the ICT and Non-ICT condition had a positive effect on the educational outcomes of pupils in the short-term. There is some evidence that effects endure, particularly in the case of the Non-ICT intervention. It would appear that the latter intervention is effective for most groups of students whereas the ICT intervention is more selective in who it benefits.

In considering mechanisms, we first discuss how to interpret differences between the treatment and control group. Then we discuss how we might interpret the spillover effect (evident in the short-term but not one year later). Finally, we discuss possible reasons for why the Non-ICT version of this intervention appears to be more effective than the ICT version.

The intended interpretation of this RCT is that differences between the treatment and control group of schools can only be attributed to the effect of training teaching assistants in the use of the pedagogy applied here. A threat to this interpretation would exist if treatment schools actually increased the hours devoted to literacy as a result of the intervention (potentially at the cost of other activities for which we have no measure of outcomes). Table 10 shows results from a survey of treatment and control schools that was undertaken at the end of the school year in which the intervention took place.[40] This shows that the hours devoted to literacy instruction was approximately the same in treatment and control schools and that schools were also similar to each other with regard to the use of computers and other forms of IT to support teaching.

Another threat to the interpretation of findings would be if there was a 'Hawthorne effect', whereby treatment schools improve relative to the control group simply because the fact of there being *any* intervention is an impetus to increase effort. This would certainly be a potential explanation for a large spillover effect within treatment schools. While one cannot rule out some effect from being put under the spotlight, the strongly heterogeneous effects of the interventions would move against such an interpretation. For example, the effects of the intervention are much stronger for pupils from disadvantaged backgrounds compared to others. This is particularly evident in the results after the first year of the intervention. Thus, the most obvious interpretation of the intervention is that the training of teaching assistants in the use of this particular pedagogy, along with its practical implementation, was effective for students.

However, the results show a strong spillover effect to control students within treatment schools. Even though this does not last beyond the year of the intervention itself, the strong magnitude of this spillover effect in the short term is something of a puzzle. A suspicion might be that the parents or teachers of students in the control condition might have found out about the methods used by the teaching assistants and started using the resources more broadly. However, the (independently conducted) process evaluation suggests that this is extremely unlikely. Firstly, it was not straightforward even to apply the intervention to the treatment groups. Logistical issues that affected the majority of TAs included taking pupils to and from sessions; space within the school and the short length of sessions. Secondly, the external process evaluation did not find that schools were compensating for the program by delivering additional help to pupils in the control group. Finally, the identity of the computer program was supressed throughout the evaluation and known only to TAs and students that saw the name of the program when actually using it.[41]

[40] The results of this exercise are informative but need to be taken with caution since the data is only available for 29 schools (out of 50 schools that were randomised).

[41] The intervention was closely monitored by the implementation team throughout (with TAs receiving visits) and fidelity to the design was strongly emphasised. TAs were asked to keep the interventions distinct by not sharing information about the content and delivery of the two programs. Process evaluators found only a low level of awareness among TAs

**Table 9**
Heterogeneous effects.

| Outcome: | PIRA at endline (1) | PIRA at endline +1 (2) | KS1 reading at endline +1 (3) |
|---|---|---|---|
| A. FSM interactions | | | |
| ICT*FSM | 0.455*** | 0.217* | 0.045 |
| | (0.136) | (0.111) | (0.059) |
| ICT*NOFSM | 0.110 | 0.043 | 0.012 |
| | (0.079) | (0.079) | (0.025) |
| NONICT*FSM | 0.482*** | 0.117 | 0.095** |
| | (0.098) | (0.091) | (0.043) |
| NONICT*NOFSM | 0.211** | 0.088 | 0.044 |
| | (0.080) | (0.092) | (0.033) |
| FSM | −0.301*** | −0.244*** | −0.086** |
| | (0.075) | (0.065) | (0.039) |
| Ho: ICT (FSM-NOFSM)=0 | 0.007 | 0.590 | 0.715 |
| Ho: NONICT (FSM-NOFSM)=0 | 0.000 | 0.357 | 0.047 |
| B. Gender interactions | | | |
| ICT*Female | 0.207** | 0.022 | 0.014 |
| | (0.089) | (0.083) | (0.028) |
| ICT*Male | 0.152* | 0.141 | 0.024 |
| | (0.089) | (0.095) | (0.039) |
| NONICT*Female | 0.341*** | 0.087 | 0.093*** |
| | (0.092) | (0.087) | (0.035) |
| NONICT*Male | 0.200** | 0.100 | 0.015 |
| | (0.091) | (0.121) | (0.042) |
| Female | −0.081* | 0.033 | −0.042 |
| | (0.045) | (0.049) | (0.032) |
| Ho: ICT (Fem-Male)=0 | 0.516 | 0.267 | 0.834 |
| Ho: NONICT (Fem-Male)=0 | 0.194 | 0.923 | 0.164 |
| C. Above/below median prior attainment (based on PIRA baseline test) | | | |
| ICT*(> median) | 0.075 | 0.043 | 0.042 |
| | (0.077) | (0.090) | (0.026) |
| ICT*(< median) | 0.278** | 0.110 | −0.004 |
| | (0.104) | (0.087) | (0.043) |
| NONICT*(> median) | 0.254*** | 0.114 | 0.054* |
| | (0.081) | (0.093) | (0.031) |
| NONICT*(< median) | 0.293** | 0.075 | 0.050 |
| | (0.114) | (0.103) | (0.050) |
| Pira baseline above median prior attainment | 0.068 | 0.047 | 0.055 |
| | (0.065) | (0.075) | (0.038) |
| Ho: ICT (Above-Below)=0 | 0.044 | 0.519 | 0.381 |
| Ho: NONICT (Above-Below)=0 | 0.767 | 0.728 | 0.946 |

Notes: Intention to treat estimates. Number of students (schools) in columns 1, 2 and 3, respectively is: 1884 (48), 1785 (48) and 2111 (50). Outcome variables: PIRA at endline is the standardised score of the PIRA test taken at the end of treatment. PIRA at endline +1 is the standardised score of the PIRA test taken a year after the end of treatment. KS1 reading at endline + 1 is a dummy variable that equals 1 if the student is at or above the expected reading level at the end of Key Stage 1. We also interact in each panel, the CT intention to treat dummy with each of the conditions explored, although we do not show the results. All regressions control for FSM and female dummy, FSP GLD, standardised baseline PIRA tests, and the randomisation pairs. Standard errors are clustered at the school level, with * p<0.10; **p<0.05; ***p<0.01.

It seems more likely that the spillover effect arises from the training to TAs, which might have affected their other activities with the Year 1 group as a whole. TAs on the project were drawn from those working with Year 1 pupils. Using data from the School Workforce Census, we calculate that TAs in Primary Schools work about 6.5 h per day on average and therefore, the intervention is estimated to have taken about 15% of their time per week (over 20 weeks). As the pupils did plenty of other literacy activities outside the intervention time, there would have been opportunity for TAs to use any new skills they had learnt to help pupils informally at other times.[42] Feedback from TAs given in the context of the process evaluation was that they perceived it to have improved their skills in small group tuition. Moreover, data from a post-treatment survey (answered by more than 70% of the TAs) shows that 74% of TAs had a better or much better understanding of phonics after the intervention, and 69% of TAs were confident or very confident to deliver small group teaching after the intervention.

Also, it is possible that the reduced number of students in the class (albeit for short periods) might have helped the class teachers with other students. Or it might be the case that the teacher was able to advance the whole class more quickly on account of the fact that two-thirds of the year group were exposed to this intervention, which complemented core literacy instruction. In any case, the spillover effect does not last into

for the training program that they were not trained to implement (in a post-treatment survey answered by 35 TAs, only 17% of the TAs answered that they saw the intervention of the other TA within their school).

[42] In general, "teaching assistances support teachers and help children with their educational and social development, both in and out of the classroom. The job will depend on the school and the age of the children". https://www.ucas.com/ucas/after-gcses/find-career-ideas/explore-jobs/job-profile/teaching-assistant

**Table 10**
A comparison between treatment and control schools, post-intervention.

| | Control Schools (1) | Treatment schools (2) | *P*-values of the difference in means [Observations] (3) |
|---|---|---|---|
| Hours of literacy instruction per week | 7.372 | 8.049 | 0.39 |
| | (1.697) | (2.790) | [48] |
| Computers are used to support literacy teaching | 0.750 | 0.726 | 0.863 |
| | (0.442) | (0.456) | [48] |
| Smartboards are used to support literacy teaching | 0.967 | 0.964 | 0.962 |
| | (0.183) | (0.190) | [48] |
| Projectors are used to support literacy teaching | 0.467 | 0.393 | 0.651 |
| | (0.509) | (0.500) | [48] |
| Tablets are used to support literacy teaching | 0.628 | 0.750 | 0.413 |
| | (0.493) | (0.443) | [48] |

Notes: The information in this table comes from data collected via surveys at endline (i.e., end of Year 1). Columns 1 and 2 show means (first row) and standard deviations (in parentheses). *P*-values are calculated using robust standard errors (column 3). [Results are very similar when we also include randomisation pairing dummies to calculate *p*-values; or when we calculate them using standard errors clustered at the school level. Due to the low number of observations and clusters, we show the results without including pairing dummies and without clustering standard errors at the school level]. Observations (i.e. number of Year 1 teachers replying to the surveys) appear in column 3 in squared brackets and have a weight of 1 if there is only one Year 1 teacher replying to the questionnaire per school; and 0.5 when there are two Year 1 teachers replying to the questionnaire per school.

**Table 11**
Compliance according to intervention type.

| | ICT (1) | Non-ICT (2) | *P*-values of the difference in means [Observations] (3) |
|---|---|---|---|
| Score based on daily record keeping by the TA (1 to 10) | 8.130 | 9.478 | 0.047 |
| | (2.916) | (1.229) | [46] |
| Score based on TA use of the levels (1 to 10) | 6.457 | 7.022 | 0.347 |
| | (2.147) | (1.880) | [46] |
| Number of weeks the TA kept records (maximum=20) | 18.28 | 19.42 | 0.158 |
| | (3.304) | (1.865) | [46] |

Notes: The information in this table comes from data collected by the implementation team. Researchers at the implementation team gave scores for daily record keeping and use of levels at the end of the implementation. Columns 1 and 2 show means (first row) and standard deviations (in parentheses). *P*-values are calculated using robust standard errors (column 3). The number of observations appears in squared brackets in column 3. Results are very similar when we also include school fixed effects or when we cluster the standard errors at the school level. Due to the low number of observations and clusters, and the fact that in the second panel we miss information for some of the TAs in some categories, we show the results without including school fixed effects and without clustering standard errors at the school level. There is only one case with two teaching assistants per group in this data. For this particular case, we consider the average score between the two teaching assistants (all the other cases have 1 observation per teaching assistant or group of teaching assistants).

the subsequent year and the Non-ICT intervention has a more enduring impact than the ICT intervention (at least on average). So why might the Non-ICT intervention have been more effective?

We first consider whether compliance was different for teaching assistants assigned to either type of intervention. Table 11 shows scores for daily record keeping and the use of levels (which indicates the extent to which TAs were moving pupils through different layers of the program adequately). These measures suggest a high level of compliance for TAs assigned to both treatments. Even though those assigned to the Non-ICT condition perform slightly better on daily record keeping, it would be hard to believe that this could explain the stronger and more enduring effect for pupils being assigned to the Non-ICT treatment. Also, although TAs were allowed to decide how to group pupils assigned to each condition, there was no difference in the size of groups or their composition between the ICT and Non-ICT condition. This is shown in Table 12.

Although one might think that technical problems could jeopardise the ICT intervention, in practice any technical problems with implementing the ICT intervention were minor and occasional. Furthermore, the process evaluation found that both interventions were extremely popular with TAs and with pupils. The training for interventions was also equally well received.[43] The process evaluation found that the Non-ICT intervention was perceived to have greater adaptability to different ability levels by TAs. This may lie at the heart of the differential effectiveness because it is consistent with the fact that the Non-ICT intervention shows stronger effects for students above and below median prior attainment (whereas the ICT intervention only shows strong effects for the latter group). Thus, it might be that when confronted with different levels of ability and progression, the TAs and pupils found it easier to use books and magnetic letters to advance learning rather than the medium of a computer screen. This is consistent with the large body of research (cited above) suggesting that computer-aided instruction is not in and of itself any better than what it replaces.[44]

This study shows that teaching assistants can be deployed very effectively to supplement classroom teaching with small, short tutorial sessions, using a highly structured evidence-based approach. Most of the TAs already had some experience of using literacy programmes with small children, but their feedback suggested that this intervention was unlike anything most had used before. The main difference was in the complete and packaged nature of the intervention and the requirement to follow it closely, including through time allocation of components within the delivery. The TAs in this study reported feeling well prepared

[43] The qualitative methods used in the process evaluation are documented in McNally et al. (2016).

[44] An additional disadvantage of the computer program in this particular context is that there were Canadian English pronunciations, which might have affected the learning experience of students.

**Table 12**
Group size and composition by treatment condition.

| | ICT (1) | Non-ICT (2) | P-values of the difference in means [Observations] (3) |
|---|---|---|---|
| Average group size | 3.597 | 3.69 | 0.35 |
| | (0.520) | (0.667) | [148] |
| *Within group standard deviations for:* | ICT | Non-ICT | *P-values of the difference in SD by group and treatment conditions* |
| FSM | 0.316 | 0.34 | 0.59 |
| Female | 0.425 | 0.426 | 0.988 |
| Standardised baseline PIRA | 0.592 | 0.566 | 0.649 |

Notes: *P*-values calculated by regressing the average group size in each small group (or the SD for each small group for the variables FSM, Female and Standardised baseline PIRA) on a dummy for the NON-ICT group, with robust standard errors. Results are very similar when we also include school fixed effects or when we cluster the standard errors at the school level. Due to the low number of observations and clusters, we show the results without including school fixed effects and without clustering standard errors at the school level. The number of observations in these regressions is 148, which corresponds to the number of small groups formed by the teaching assistants overall (i.e., in both ICT and NON-ICT conditions). There is no information on the groups for the 3 schools in the treatment group that dropped out immediately after randomisation.

for the intervention in terms of training and well supported throughout by the implementation team.

## 7. Conclusion

In this study, we get inside the 'black-box' of the education production function from within the classroom. The experiment provides an opportunity to evaluate whether teaching assistants can be effectively deployed to complement the work of the teacher. This study shows a context of how teaching assistants (who are employed by almost all primary schools in England) can be used to better effect to improve the literacy of young children. Teaching training has been shown to be important in other contexts (e.g. Angrist and Lavy, 2001). Here we show that training of teaching assistants can also be an effective way to improve student outcomes.

Further, we are able to distinguish the effects of the training of TAs and pedagogy from the effect of the medium of delivery of the intervention (whether ICT or Non-ICT). Although both modes of delivery show positive effects on pupil outcomes, the Non-ICT mode of delivery has a stronger and more enduring effect. This shows that although computer-aided instruction can be useful, it does not (in and of itself) add value to such pedagogical approaches.

Given that both interventions were delivered by TAs already employed by the schools, who are not very highly qualified (or highly paid), the per-pupil costs of delivering this intervention were modest. We estimated that the per-pupil cost (including the training of TAs; support provided during the project etc.) was about £25. This assumes that existing TAs and computers can be used for project implementation.[45] This low per pupil cost implies that effects do not have to be very large before the intervention becomes cost effective. Although there is some evidence of fade-out, the one year follow up does suggest that effects endure (at least beyond the year of the intervention). This is most evident with respect to the effect of the Non-ICT intervention on the probability of being at or above the 'expected level' at age 7 in teacher assessments of reading and writing.

Finally, this is an intervention that disproportionately benefits students from a lower socio-economic background. Although this is most evident for short-term outcomes, it is also true for outcomes measured one year later. Thus, using teaching assistants effectively in the context of an intervention such as this one helps to level the playing field between pupils from different socio-economic groups.

[45] This was the case in this study. For this study, laptops were supplied to TAs. However, most primary schools in England are well-equipped with ICT and all employ TAs.

## Acknowledgements

We would like to thank the Education Endowment Fund (EEF) for commissioning this research, which was implemented by the research team at Coventry University (involving Johnson, Vousden, Savage, Wood and their research teams) and independently evaluated by the team at the Centre for Economic Performance, LSE (involving Ruiz–Valenzuela and McNally) and the National Institute of Economic and Social Research (Rolfe). We thank the team of research assistants at Coventry University that facilitated the fieldwork, the technical support provided by Annie Wade and Philip Abrami, and others at Concordia's Centre for the Study of Learning and Performance (CSLP). We also thank participating schools and teaching assistants. We thank Robert Slavin for evaluating the training materials and Morag Stuart for implementing the Non-ICT version of the program evaluated here. We thank participants at seminars at Trondheim University, the Centre for Economic Performance (London School of Economics); and at the CESifo Area Conference on Economics of Education, the Royal Economic Society conference, the II Workshop on Empirical Research in Economics of Education (Universitat Rovira i Virgili) and the 30th conference of the European Association of Labour Economists.

## Appendix

*Note on Methodology*

The first stages for whether students are in the final ICT or final Non-ICT treatments, or in the final CT group (i.e. control students in treatment schools) are as follows:

$$ICT\ Final_{ist} = \gamma_1 ICT_{ist} + \gamma_2 NonICT_{ist} + \gamma_3 CT_{ist} + \gamma_4 Y_{ist-1} + \gamma_5 X_{ist-1} + \rho_s + \varepsilon_{ist} \quad (A1)$$

$$NonICT\ Final_{ist} = \pi_1 ICT_{ist} + \pi_2 NonICT_{ist} + \pi_3 CT_{ist} + \pi_4 Y_{ist-1} + \pi_5 X_{ist-1} + \rho_s + \varepsilon_{ist} \quad (A2)$$

$$CT\ Final_{ist} = \beta_1 ICT_{ist} + \beta_2 NonICT_{ist} + \beta_3 CT_{ist} + \beta_4 Y_{ist-1} + \beta_5 X_{ist-1} + \rho_s + \varepsilon_{ist} \quad (A3)$$

Where $ICT\,Final_{ist}$ ($NonICT\,Final_{ist}$) is a dummy variable equal to 1 if students received the *complete* 20-week ICT (Non-ICT) intervention, and equal to 0 otherwise. $CT\,Final_{ist}$ is a dummy variable equal to 1 if

**Table A1**
Attrition.

| | Control Schools (1) | Treatment schools (2) | ICT (3) | Non-ICT (4) | Control in Treatment schools (5) |
|---|---|---|---|---|---|
| Students initially allocated to… | 1158 | 1083 | 360 | 350 | 373 |
| *Fraction students in each group with….* | | | | | |
| Missing baseline PIRA | 0.030 | 0.024 | 0.033 | 0.020 | 0.019 |
| Missing endline PIRA | 0.047 | 0.153 | 0.150 | 0.171 | 0.139 |
| Missing endline Key Stage 1 Reading at t + 1 | 0.020 | 0.028 | 0.028 | 0.034 | 0.021 |
| Missing endline PIRA at t + 1 | 0.108 | 0.189 | 0.186 | 0.211 | 0.172 |

*Note.* Key Stage 1 data is available for all schools that were included in the randomisation. Five schools in the treatment group dropped out after randomisation (3 right after randomisation, 2 during the intervention). Post-intervention tests right at the end of the intervention and at t + 1 were conducted in all schools but 1.

**Table A2**
IV estimates.

| A. Outcome: | PIRA at endline (1) | PIRA at endline (2) | PIRA at endline+1 (3) | PIRA at endline+1 (4) | KS1 read at endline+1 (5) | KS1 read at endline+1 (6) |
|---|---|---|---|---|---|---|
| ICT | 0.172* | 0.216** | 0.063 | 0.092 | 0.011 | 0.024 |
| | (0.103) | (0.092) | (0.086) | (0.083) | (0.032) | (0.032) |
| NONICT | 0.297*** | 0.328*** | 0.086 | 0.113 | 0.064* | 0.073** |
| | (0.099) | (0.088) | (0.098) | (0.091) | (0.035) | (0.034) |
| CT | 0.139 | 0.201** | −0.025 | 0.019 | −0.028 | −0.009 |
| | (0.097) | (0.088) | (0.092) | (0.086) | (0.031) | (0.032) |
| B. Main coefficient in ICT first stage | | | | | | |
| Randomised to ICT | 0.845*** | 0.843*** | 0.844*** | 0.843*** | 0.759*** | 0.758*** |
| | (0.067) | (0.068) | (0.067) | (0.067) | (0.090) | (0.090) |
| F-test of excluded instruments | 84.070 | 72.940 | 73.340 | 71.470 | 45.510 | 44.850 |
| C. Main coefficient in NON-ICT first stage | | | | | | |
| Randomised to NONICT | 0.829*** | 0.831*** | 0.835*** | 0.835*** | 0.749*** | 0.751*** |
| | (0.072) | (0.070) | (0.070) | (0.069) | (0.092) | (0.091) |
| F-test of excluded instruments | 59.660 | 60.830 | 69.340 | 70.420 | 39.810 | 43.010 |
| D. Main coefficient in CT first stage | | | | | | |
| Randomised to NONICT | 0.849*** | 0.847*** | 0.842*** | 0.840*** | 0.770*** | 0.771*** |
| | (0.064) | (0.064) | (0.066) | (0.066) | (0.086) | (0.086) |
| F-test of excluded instruments | 92.760 | 83.280 | 76.990 | 73.860 | 49.700 | 49.000 |
| Students | 1901 | 1884 | 1799 | 1785 | 2129 | 2111 |
| Schools | 48 | 48 | 48 | 48 | 50 | 50 |
| Baseline PIRA test | √ | √ | √ | √ | √ | √ |
| Gender, FSM, FSP GLD | | √ | | √ | | √ |

Notes: Instrumental variable estimates. Outcome variables: PIRA at endline is the standardised score of the PIRA test taken at the end of treatment. PIRA at endline +1 is the standardised score of the PIRA test taken a year after the end of treatment. KS1 reading at endline + 1 is a dummy variable that equals 1 if the student is at or above the expected reading level at the end of Key Stage 1. ICT and NONICT are the endogenous treatment dummies. CT is the endogenous treatment dummy equal to 1 for pupils in the control group of treatment schools as their final assignment. All regressions control for the randomisation pairs. Standard errors are clustered at the school level, with * p<0.10; **p<0.05; ***p<0.01.

students were in the control group of treated schools that implemented the 20-week programs. The second stage equation is then given by:

$$Y_{ist} = \theta_1 ICT\ Final_{ist} + \theta_2 NonICT\ Final_{ist} + \theta_3 CT\ Final_{ist} + \theta_4 Y_{ist-1} + \theta_5 X_{ist-1} + \rho_s + \varepsilon_{ist} \quad \text{(A4)}$$

We estimate (A4) by two stage least squares, using the initial random allocations, $ICT_{ist}$, $NonICT_{ist}$ and $CT_{ist}$, respectively, as instruments for $ICT\ Final_{ist}$, $NonICT\ Final_{ist}$ and $CT\ Final_{ist}$ and the other variables as instruments for themselves.

## References


Aaronson, D., Barrow, L., Sander, W., 2007. Teachers and student achievement in the chicago public high schools. J. Labor Econ. 25, 95–135.

Abrami, P.C., Savage, R.S., Deleveaux, G., Wade, A., Meyer, E., Lebel, C., 2010. The Learning Toolkit: the design, development, testing and dissemination of evidence-based educational software. In: Zemliansky, P., Wilcox, D.M. (Eds.), Design and Implementation of Educational Games: Theoretical and Practical Perspectives. IGI Global, Hershey, PA, pp. 168–187. http://dx.doi.org/10.4018/978-1-61520-781-7.ch012.

Angrist, J., Lavy, V., 2001. Does teacher training affect pupil learning? evidence from matched comparisons in Jerusalem public schools. J. Labour Econ. 19 (2), 343–369.

Angrist, J., Lavy, V., 2002. New evidence on classroom computers and pupil learning. Econ. J. 112, 735–765.

Araujo, M., Carneiro, P., Cruz-Aguayo, Y., Schady, N., 2016. Teacher quality and learning outcomes in kindergarten. Q. J. Econ. 131, 1415–1453.

Archer, K., Savage, R.S., Sanghera-Sidhu, S., Wood, E., Gottardo, A., Chen, V., 2014. Examining the effectiveness of technology use in classrooms: a tertiary meta-analysis. Comput. Edu. 78, 140–149.

Berlinski, S., Busso, M., 2017. Challenges in educational reform: an Experiment on active learning in mathematics. Econ. Lett. 156, 172–175.

Bulman, R., Fairlie, R.W., 2016. Technology and education: computers, software and the Internet. In: Hanushek, E.A., Machin, S.J., Woessmann, L. (Eds.). In: Handbook of the Economics of Education, 5, pp. 239–280.

Chetty, R., Friedman, J.N., Rockoff, J., 2014a. Measuring the impacts of teachers I: evaluating bias in teacher value-added estimates. Am. Econ. Rev. 104, 2593–2632.

Chetty, R., Friedman, J.N., Rockoff, J., 2014b. Measuring the impacts of teachers II: teacher value-added and student outcomes in adulthood. Am. Econ. Rev. 104, 2633–2679.

Comaskey, E.M., Savage, R.S., Abrami, P.C., 2009. A randomised efficacy study of Web-based synthetic and analytic programmes among disadvantaged urban kindergarten children. J. Res. Reading 32, 92–108.

Hanushek, E., Rivkin, S., Kain, J., 2005. Teachers, schools and academic achievement. Econometrica 73, 415–458.
Kuczera, M., Field, S., Windisch, H., 2016. Building Skills for All: A Review of England. Policy Insights from the Survey of Adult Skills. OECD Skills Studies.
Machin, S., McNally, S., 2008. The literacy hour. J. Public Econ. 92, 1441–1462.
Machin, S., McNally, S., Viarengo, M., 2018. Changing how literacy is taught: evidence on synthetic phonics. Am. Econ. J. 10 (2), 217–241.
McNally, S., Rolfe, H., Ruiz-Valenzuela, J., 2016. ABRA: Online Reading Support Evaluation report and executive summary. Report for the Education Endowment Foundation https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/EEF_Project_Report_ABRA.pdf.
Rouse, C., Krueger, A., Markman, L., 2004. Computerized Instruction to the Test. NBER, p. 10315.
Savage, R.S., Abrami, P., Hipps, G., Deault, L., 2009. A randomized controlled trial study of the ABRACADABRA reading intervention program in grade 1. J. Edu. Psychol. 101 (3), 590.
Savage, R., Abrami, P.C., Piquette, N., Wood, E., Deleveaux, G., Sanghera-Sidhu, S., Burgos, G., 2013. "A (Pan-Canadian) cluster randomized control effectiveness trial of the ABRACADABRA web-based literacy program. J. Edu. Psychol. 105 (2), 310.
Slavin, R.E., Lake, C., Davis, S., Madden, N.A., 2011. Effective programs for struggling readers: a Best-evidence Synthesis. Edu. Res. Rev. 6, 1–26.
Spybrook, J., Bloom, H., Congdon, R., Hill, C., Martinez, A., Raudenbush, S.. Optimal design plus empirical evidence: documentation for the "Optimal Design" software.
Taylor, E.S., Tyler, J.H., 2012. The effect of evaluation on teacher performance. Am. Econ. Rev. 102 (7), 3628–3651.
Vignoles, A., 2016. What is the Economic Value of Literacy and Numeracy? Basic Skills in Literacy and Numeracy are Essential for Success in the Labor Market. IZA World of Labor, p. 229.
Wolgemuth, J.R., Savage, R.S., Helmer, J., Lea, T., Harper, H., Chalkiti, K., Bottrell, C., Abrami, P., 2011. Using computer-based instruction to improve indigenous early literacy in Northern Australia: a quasi-experimental study. Australas. J. Edu. Technol. 27, 727–750.
Whitmore Schanzenbach, D., 2007. What have researchers learned from Project Star? Brookings Papers Edu. Policy 9 (2006/2007), 205–228.